International Journal of Scientific Research and Engineering Development-– Volume 4 Issue 3, May -June 2021

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Resolution in Ambiguity in Machine Translation Using Natural Language Processing Techniques

Pankaj V. Nimbalkar

*(Department of Computer Science, Dr.Ambedkar College, Deekshabhoomi, Nagpur-10,India

Email:pankajnimbalkar3@gmail.com)

Abstract:

Machine translation (MT) plays a vital role in translation of text from one language to another corpus-based, and knowledge-based. It also highlights characteristics Indian languages and translation strategies. Systems of Translation that automatically take out transfer mappings (rules or examples) from bilingual corpora have been held back by the difficulty to achieve accurate alignment and acquire high quality mappings.[1]

Keywords: Word sense disambiguation (WSD); Natural Language Processing (NLP)

1 INTRODUCTION

Machine translation (MT), perhaps the earliest NLP application is translation of text units from one language into another, using computers. It is fascinating to think of an MT system that can translate literary from any language into our native language. So we need not to know Urdu to read the stories of Manto; just feed it to an MT system and get translated. Such MT systems are able to break the barriers of languages by making available rich sources of literature to people across the world. A machine translation system requires a significant amount of translation facts typically personified in bilingual dictionaries, transfer rules, example bases, or a statistical model.[2]

1.1NLP flow chart: Word sense disambiguation permits the user to introduce rectifications into the decisions made by the flow chart. In the first appearance the user types in the label and launches processing by making no input during a already defined period of time. In the second appearance the user can correct the results before the final translation into the logical formula. The flow chart API allows manipulating and correcting the results of the automatic processing.

1.2 Text Sense Disambiguation

The unit of a natural language is a text, which is often ambiguous. To translate natural language into formal concept language, disambiguate ambiguous texts into unambiguous concepts. Word Sense Disambiguation is a standard problem

in natural language processing. Traditional texts sense disambiguation algorithms developed with a model .Natural language data dictionary is widely used in various applications.The natural language metadata is ambiguous and proposed solution assists in tackling the problem asises. In other words, the best place and time to get liberate of the ambiguity of a piece of a natural language metadata is when it is created and where it is created. At that time the user who creates it is right there and, seeing this as part of the creation process is more willing to cooperate in the task of creating a proper semantic annotation, on the other hand to returning later to the data and doing the footnote from scratch. This leads to applications which might benefit if they integrate the proposed solution on the user interface level, and solve the problem in cooperation with the user. Many texts have more than one meaning; we have to select the meaning which makes the most sense in context. For this problem, we can find out a list of texts and associated text senses, e.g. from a dictionary or from an online resource such as Word Net[3].

2. PROBLEMS IN MACHINE TRANSLATION

There are many structural and stylistic differences among languages which make automatic translation a difficult task.

2.1 Word Order

The arrangement of words in a sentence varies across languages. For example, in English words are usually arranged in the order subject, verb and object; whereas in Indian languages, object usually precedes verb. This makes a word by word translation impractical.

2.2Word Sense

The sense of a word in one language may translate into a different sense with the words of another language. This creates problem in target language word selection.

2.3Pronoun Resolution

Resolving pronomial references is important for machine translation. Unresolved references may lead to incorrect translation.

2.4 Idioms

A sentence involving natural expressions is difficult to translate as idioms are composed of words that do not directly contribute to their meaning. Replacing words constituting an idiom with words from the target language can lead to funny and nonsensical translations. For example, consider the sentence: ‘The old man finally kicked the bucket.’ If the system does not. recognize the idiom ‘kicked the bucket’ the translation, say in Hindi, will end up as:

‘Boodhe aadmi ne ant-ta batli me laat mari.

3.Ambiguity

Certain languages do not permit certain types of ambiguities. For example

consider the PP-ambiguity in the English sentence: 'The man saw the girl with a telescope.' In order to translate this sentence into Hindi PP-ambiguity must first be resolved.

Ambiguity can occur in four different level

3.1 Lexical Ambiguity:-Lexical ambiguity is the ambiguity of a single word. A word can be ambiguous with respect to its internal structure or to its syntactic class. Lexical semantic ambiguity occurs when a single word is associated with multiple senses. This type of ambiguity is viewed as a part of speech tagging in NLP and has been solved with reasonable accuracy.

Syntactic Ambiguity:-The structural ambiguity is syntactic ambiguity in which meaning of word is not clear and depend upon preposition

For Ex:-The man saw a girl with Telescope

The above sentence is ambiguous whether the man saw a girl carrying telescope or saw her through telescope. This ambiguity is also called as PP attachment ambiguity.

3.2 Semantic Ambiguity:-In semantic ambiguity meaning of words themselves can be misinterpreted. The meaning of words in a phrase can be combined in different words

Pragmatic Ambiguity:-In Pragmatic ambiguity the context of phrase gives multiple interpretation .This type of ambiguity requires discourse processing.

4.POS Tagging

In order to find out whether or not morphosyntactic ambiguity could provide useful information to recognize a valid sentence, carried out an experiment to know if the number of POS tags that a word could have in a sentence could be a feature to distinguish the morphosyntactic ambiguity of our sets. The process we performed was to label both positive and negative sets using the TreeTagger software (Schmid, 1995). The only requirement was to label the sets with a threshold of probability of (0,3) in order to get all the possible tags that a word could have. The number of tokens for both sets are:171,327 for the positive set and 228,522 for the negative one. The total tags that could be assigned to every token is shown in Table .

We can see in this table how, considering the number of tokens per set, the OLs have a wider range of possibilities to be tagged. This information strengthens the previous behaviour that the OLs follow a pattern of ambiguity in order The Impact of Semantic and Morphosyntactic Ambiguity.[5]

Table 1 : Number of tags for OLs and LC

Tags	Positive Set	Negative Set
2	13,926	13,450
3	1,679	1,427
4	142	65
5	35	4

6 1 0

to produce humour. Thus, based on the information of this table, we can consider that, given a word wi from the OLs set, there is more ambiguity in the morphosyntactic functions that a word can play (due to the fact that wi may be noun, verb, adjective, etc., with higher probability than a word wj from the LC). Therefore, this behaviour may be a feature that generate ambiguity and consequently, a potential comical situation.

Table 2: Number of words per category

Category	Positive Set	NegativeSet
N	3,466	3,300
ADJ	926	1,172
ADV	265	349

5. Discussion

According to Mihalcea and Strapparava (2006a, 2006b) ambiguity is one of the sources which potentially produces humour.Therefore, using some illustrations whose final goal was to analyse whether it could be possible to find some features in order to characterize the

humour of the OLs taking into account linguistic information. On the basis of these premises, the results of each illustration suggest the following inferences.

The perplexity reported in Table 1 indicates that words of the Positive set have a wider range of combination. That is, in terms of language models, it is more probable to predict the word w+1 given a word w which belongs to the negative sets

. This information points out a minor predictability in the positive set. Thus, on the basis of this fact, we think that this kind of dispersion is a focus of ambiguity that may produce humour.

The results of the POS tagging experiment have shown how the words of the positive set have a bigger probability to play different morphosyntactic functions.

In Table 2 it can be observed how this diffusion that we assume as ambiguity is greater when a tag was assigned to a positive word than when it was assigned to a negative one. On the basis of this behaviour, we may conclude that, given a remote word wi, the probability to be assigned to various categories can break a logical meaning producing ambiguity and, therefore, a humorous effects.

The results we obtained suggest that the sentences of LC are syntactically more complex than the OLs. This behaviour supports what Mihalcea and

of the OLs, and rejects our hypothesis about the SC as a feature to characterize the syntactic ambiguity. Therefore, on the basis of this result, we can consider that the ambiguity does not appear from a syntactic viewpoint because the Ols are well formed structures that utilize other kind of strategies, (for instance,semantic and pragmatic information) in order to produce humour.

The impact of the semantic information as a trigger to generate humour seems to be more relevant. There are more elements that strengthen the assumption about that humour is produced employing semantic information that generates ambiguity.

5.Conclusions and Further Work

In this paper we try to investigate one of the most important sources that generate humorous situations:ambiguity. We analysed, from different linguistic layers,whether the information that we obtained studying the ambiguity could be taken into account as a set of features to automatically recognize humour or not. Some of the results obtained confirmed our initial assumptions about the usefulness of this kind of information to characterize humorous examples, especially with respect to measures such as perplexity, mean of senses and sense dispersion. As final remark, we can say that the OLs contain elements that could potentially be considered as triggers of humour, i.e., the OLs are ambiguous but, according to our results, not from a syntactic viewpoint. This behaviour could be due to the fact that the OLs are syntactically well formed structures which exploit ambiguous referents related to words and not to the whole sentence. Summarizing,although the preliminary findings are interesting, they must be further tested with a classifier. We also plan, as further work, to integrate these features in a mixture model in order to identify their relevance.

6. References

[1] Mani, I., Pustejovsky, J., Gaizauskas, R.: The Language of Time: A Reader. Oxford University Press, Oxford (2005)

[2] Schilder, F., Katz, G., Pustejovsky, J.: Annotating, Extracting and Reasoning About Time and Events. In: Schilder, F., Katz, G., Pustejovsky, J. (eds.) Annotating, Extracting and Reasoning about Time and Events. LNCS (LNAI), vol. 4795, pp. 1–6. Springer, Heidelberg (2007)

[3]. Mani, I., Wilson, G.: Robust temporal processing of news. In: ACL AnnualMeeting, NJ, USA, ACL, pp. 69–76 (2000)

[4] Pustejovsky, J.: TERQAS: Time and Event Recognition for Question Answering Systems. In: ARDA Workshop (2002)

[5] Pustejovsky, J., Mani, I., Belanger, L., Boguraev, B., Knippen, B., Litman, J., Rumshisky, A., See, A., Symonen, S., van Guilder, J., van Guilder, L., Verhagen, M.: ARDA summer workshop on graphical annotation toolkit for TimeML. Technical report, MITRE (2003)

[6] TERN-2004: Time Expression Recognition and Normalization Evaluation Workshop (2004), http://fofoca.mitre.org/tern.html

[7] Verhagen, M., Gaizauskas, R.J., Hepple, M., Schilder, F., Katz, G., Pustejovsky, J.: Semeval-2007 task 15: Tempeval temporal relation identification. In: Proceedings of the 4th InternationalWorkshop on Semantic Evaluations, ACL, pp. 75–80 (2007)

[8] Setzer, A., Gaizauskas, R.: Annotating Events and Temporal Information in Newswire Texts. In: LREC 2000, Athens, pp. 1287–1294 (2000)

[9] Ferro, L., Gerber, L., Mani, I., Sundheim, B., Wilson, G.: TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, MITRE (2005)

[10] Pustejovsky, J., Casta˜no, J.M., Ingria, R., Saur´ı, R., Gaizauskas, R.J., Setzer, A., Katz, G.: TimeML: Robust Specification of Event and Temporal Expressions in Text. In: IWCS-5, 5th Int. Workshop on Computational Semantics (2003)

